Enfoque probabilístico para la construcción de ontologías

Isidra Ocampo-Guzman¹, Ivan Lopez-Arevalo¹, Edgar Tello-Leal²,
Victor Sosa-Sosa¹

¹Laboratorio de Tecnologías de Información
Cinvestav - Tamaulipas. Victoria, Tamaulipas, Mexico

²Universidad Autónoma de Tamaulipas.
Victoria, Tamaulipas, Mexico

{iocampo, ilopez, vsosa}@tamps.cinvestav.mx, etello@uat.edu.mx

Abstract. *Este artículo propone una metodología para la construcción automática de ontologías a partir de un corpus de texto. Se identifican los temas de los documentos del corpus mediante Latent Dirichlet Allocation (LDA). Con base en el conjunto de temas identificados se construye para cada uno de éstos su taxonomía con sus términos de mayor probabilidad, utilizando el tesoro WordNet, mediante el cual se obtiene la similaridad y relación entre los términos. Las taxonomías resultantes se unen para formar la ontología final.*

1. Introducción

Las ontologías tienen un papel clave para estructurar, manejar, compartir y procesar la información en la Web Semántica, corpus digitales, comercio electrónico, aplicaciones médicas, etc. ya que proveen conocimiento compartido y común de un determinado dominio [1].

Los modelos probabilísticos han sido utilizados para la construcción de ontologías, tal como el trabajo de Jian-hua Yeh [2] que utiliza LDA para extraer los temas, con los cuales se construye la ontología aplicando el algoritmo de clustering aglomerativo jerárquico. Zavitsanos et al. [3] aplican LDA de manera repetida con diferente número de temas. Los temas aprendidos son directamente usados como conceptos que forman la estructura de la ontología. Fortuna et al. [4] utilizan el modelo TF/IDF [5] y Latent Semantic Indexing (LSI), para asignar términos a un tema, haciendo cada tema una distribución sobre términos. Otro de los enfoques en la construcción de ontologías es utilizar técnicas de Procesamiento del Lenguaje Natural, por mencionar el trabajo de Cimiano et. al [6], basado en la hipótesis distribucional extraen un conjunto de relaciones verbo/argumento como entrada al modelo de clustering *Análisis Formal de Conceptos* para la construcción de ontologías.

En el enfoque propuesto en este artículo se construye la ontología identificando un conjunto de temas mediante el modelo Latent Dirichlet Allocation (LDA) [7]. Se construye para cada uno de ellos su taxonomía con sus términos de mayor probabilidad. Se utiliza el tesoro de WordNet [8] para identificar el término de mayor similaridad¹ con el resto para ser la raíz de la taxonomía y la relación² semántica entre los términos que forman al tema. Para determinar la similaridad y relación semántica se asocia cada

¹Cuantifica qué tanto dos conceptos son iguales con base en la información contenida en WordNet.

²Indica los sentidos en que están relacionados dos conceptos, es información no jerárquica.

término del tema con un número de glosa³ de WordNet, la cual se determina aplicando TF-IDF del modelo espacio vectorial (MEV) [5]. Finalmente, se unen las ontologías cuya similaridad de sus nodos raíces sea mayor que un determinado *umbral*.

El resto de este artículo se encuentra estructurado de la siguiente manera. En la sección 2 se presenta una breve descripción del modelo utilizado. En la sección 3 se describe la metodología propuesta. La sección 4 contiene los experimentos, resultados y evaluación. Finalmente, en la sección 5 se presentan algunas conclusiones.

2. Latent Dirichlet Allocation

Latent Dirichlet Allocation [7] es un modelo temático probabilístico que captura las características *significativas* de los documentos del corpus. Consiste en un proceso generativo en donde los documentos de un corpus son generados como una mezcla temas (distribución multinomial de probabilidad sobre temas), donde los temas están formados por una mezcla de términos del vocabulario del corpus (distribución multinomial de probabilidad sobre el vocabulario). El proceso generativo para cada documento dado un número de temas K , es el siguiente:

1. Seleccionar $N \sim \text{Poisson}(\varepsilon)$
2. Seleccionar $\theta \sim \text{Dirichlet}(\alpha)$
3. Para cada uno de los N términos w_n :
 - Seleccionar un tema $z_n \sim \text{Multinomial}(\theta)$
 - Seleccionar un término w_n de la distribución de probabilidad multinomial del tema z_n definida por $p(w_n|z_n, \beta)$.

$p(z_n=i)$ representa la probabilidad que el i^{th} tema sea muestreado para los n términos e indica los temas que son importantes (que reflejan el contenido) para un documento particular. Por su parte $p(w_n|z_n=i)$ representa la probabilidad de ocurrencia de un término w_n dado un tema i indica la probabilidad de ocurrencia de un término para cada tema.

3. Metodología

La metodología propuesta para construir la ontología comprende las siguientes fases:

1. *Pre-procesamiento*: para identificar los sustantivos candidatos que podrían formar la ontología, se etiquetan las palabras de cada sentencia de los documentos para obtener su categoría gramatical. Se eliminan todas aquellas palabras que no sean sustantivos.
2. *Creación de la matriz de características*: se crea una matriz de documentos-palabras, donde cada vector representa un documento y las columnas a los términos del vocabulario. La matriz resultante tendrá un número de filas equivalente al número de documentos y un número de columnas equivalente al total de términos en el vocabulario. La matriz es la entrada del modelo LDA.
3. *Extracción de temas*: Se infiere un conjunto T de temas del corpus mediante la aplicación del proceso generativo inverso de los modelos probabilísticos. Es decir, a partir del conjunto de documentos ya conocidos se identifican los *temas* abordados por éstos aplicándoles el modelo LDA. La inferencia exacta de los temas es muy costosa, por lo que, se construye una cadena de Markov con técnicas de Monte Carlo utilizando un muestreador que sigue el modelo LDA.

³Descripción de un *synset* (verbo, sustantivo, adjetivo o adverbio) contenido en WordNet.

4. *Construcción de la taxonomía de cada tema*: Para cada tema i del conjunto de temas T se construye su taxonomía en tres fases:

- *Extracción de términos*: los temas están compuestos por una mezcla de probabilidades sobre el vocabulario del corpus. La taxonomía de cada tema se construye utilizando un subconjunto de términos del vocabulario: para cada tema i del conjunto de temas T se extraen aquellos términos cuya probabilidad sea mayor que la mediana de su distribución de probabilidad T_i .

- *Extracción del sentido de los términos*: para la construcción de la taxonomía de un tema se obtiene la similaridad entre los términos seleccionados en la fase anterior, por lo que se asocia a cada término w_i un sentido de éste en WordNet, el cual se determina utilizando el modelo espacio vectorial (MEV) [5]. Se obtienen todos los sentidos en WordNet del término en cuestión. Cada sentido se considera como un documento, el cual se modela como un vector de acuerdo al MVE. Se construye una *consulta* con los términos seleccionados del tema al que corresponde el término y se aplica la medida de similaridad de cosenos para obtener el número de *glosa* más similar a la consulta, que se asociará al término w_i .

- *Construcción de taxonomía de cada tema*: se calcula la similaridad y la relación semántica entre los términos seleccionados de cada tema, las cuales se obtienen mediante las medidas Gloss Vector⁴ y Wu and Palmer⁵ (WUP). El proceso es el siguiente:

(a) La raíz de la taxonomía es el término i cuya sumatoria de similaridad sea mayor con respecto a los demás términos del tema. Para cada término i se obtienen las combinaciones de pares de términos que contienen a i y otro término del resto que conforman al tema. Para cada combinación se obtiene la similaridad Gloss Vector, la cual se multiplica por la probabilidad del término i en el tema.

(b) Se agregan a la taxonomía aquellos términos cuya profundidad obtenida con la medida de relación WUP sea mayor que la del término raíz y su similaridad supere un determinado *umbral*. Para cada término a agregar se compara con los términos contenidos hasta el momento en la taxonomía, se agrega el término como nodo hijo del nodo con el cual tenga mayor similaridad.

Este proceso se repite para cada tema i del conjunto T y se obtiene un conjunto J de taxonomías.

5. *Unión de taxonomías*: el conjunto de taxonomías J se une para obtener la ontología final del corpus. Se obtiene la similaridad entre pares de raíces del conjunto J de taxonomías, se unen aquellas que sobrepasen un *umbral* establecido. Se valida la profundidad de los términos raíces considerando dos casos: - *Profundidad de taxonomías iguales*: para unir las dos taxonomías se obtiene el concepto en WordNet que subsume a los términos raíces. La taxonomía resultante se agrega al conjunto J de taxonomías y se eliminan las dos taxonomías semillas. - *Profundidad de taxonomías diferentes*: se identifica la taxonomía de menor profundidad

⁴Medida de similaridad que forma vectores de segundo orden de co-ocurrencia de términos utilizando las definiciones de los conceptos de WordNet. Obtiene el coseno del ángulo entre las glosas de los conceptos para determinar la similaridad entre dos conceptos.

⁵Medida de relación que calcula la relación de dos conceptos considerando la profundidad de éstos en la jerarquía de WordNet.

(*TG*) y la de mayor profundidad (*TE*). Se considera a *TG* más general con respecto a *TE*, se agrega *TE* a *TG* en el nivel *K*, el cual es igual a la profundidad de *TE*. Si el conjunto de nodos en *TG* en el nivel *K-1* es mayor que 1, *TE* se agrega al nodo con mayor similaridad en dicho conjunto. Se elimina *TE* del conjunto *J* de taxonomías.

El proceso anterior se repite mientras existan pares de taxonomías en el conjunto *J* que superen el *umbral*. Finalmente se unen las taxonomías del conjunto *J* bajo un nodo raíz *ROOT*.

4. Resultados preliminares y Evaluación

4.1. Experimentos

La metodología propuesta se evaluó empleando el corpus Lonely Planet. La ontología de este corpus construida manualmente contiene 96 conceptos extraídos de 1801 documentos de texto del dominio de turismo.

El etiquetado del corpus se realizó con la herramienta Stanford POS Tagger. Para la extracción de los temas del corpus se utilizó el muestreador Gibbs⁶. Los valores de los hiperparámetros de LDA α y β utilizados para el muestreo son de $50/K$ y 0.01 respectivamente. Se realizaron 31 experimentos para el total de temas *K* a obtener del corpus, los mejores resultados se obtuvieron con un valor de $K=25$ y 10,000 iteraciones de muestreo. Se desarrollaron diversas pruebas para determinar el valor del *umbral* de *similaridad* entre términos y éste se estableció en 0.7 ya que arrojó mejores resultados.

4.2. Resultados y Evaluación

Para evaluar los resultados obtenidos se utiliza la medida *F* que combina las métricas de precisión y cobertura. La precisión se refiere a la proporción de conceptos correctamente identificados con respecto al total de conceptos identificados en la ontología construida y la cobertura se refiere a la proporción de conceptos identificados correctamente con respecto al número de conceptos de la ontología de referencia.:

$$F = \frac{2 * precision * cobertura}{precision + cobertura} \quad (1)$$

Se identificaron 115 conceptos de los cuales 61 corresponden a los 96 contenidos en la ontología contruida de manera manual. Los resultados presentados en la tabla 1 se comparan con el trabajo de Cimiano et. al [6] y Zavitsanos et al.[3], omitiendo en esta comparación a los trabajos de Jian-hua Yeh [2] y Fortuna et al. [4] que aunque utilizan modelos probabilísticos no evalúan sus resultados utilizando la medida *F*.

Las tres metodologías fueron evaluadas en diferentes corpus de texto. Los mejores resultados de acuerdo a la medida *F* son los obtenidos por el modelo de Zavitsanos et al., sin embargo ellos no realizan el etiquetado de la estructura de la ontología, el cual infieren mediante el traslape con la estructura de la ontología de referencia del corpus.

⁶<http://gibbslda.sourceforge.net/>.

Conceptos identificados			
Metodología	Precisión	Cobertura	Medida F
Zavitsanos	74%	85%	79%
Cimiano	29.33%	65.49%	44.69%
<i>Propuesta</i>	53%	63%	58%

Table 1. Resultados

5. Conclusiones

Este artículo propone un enfoque para construir ontologías a partir de un corpus aplicando el modelo probabilístico LDA, el cual permite identificar los temas abordados por los documentos del corpus. Se parte de la idea de que cada tema trata sobre un determinado concepto y que los términos que lo componen deben de tener una taxonomía que describa a éste. Se utiliza el tesoro WordNet para determinar la similaridad y relación semántica entre términos que conforman a cada tema. Aunque WordNet es fundamental en la propuesta presentada, puede verse también como una desventaja ya que la construcción de la taxonomía de cada tema depende totalmente de ella. En el corto plazo se experimentará la utilización de técnicas de Procesamiento del Lenguaje Natural que permitan obtener la similaridad entre términos.

References

- [1] Thomas R. Gruber. Toward principles for the design of ontologies used for knowledge sharing. *Int. J. Hum.-Comput. Stud.*, 43(5-6):907–928, 1995.
- [2] Jian-Hua Yeh and Naomi Yang. Ontology construction based on latent topic extraction in a digital library. In *ICADL 08: Proceedings of the 11th International Conference on Asian Digital Libraries*, pages 93–103, Berlin, Heidelberg, 2008. Springer-Verlag.
- [3] Elias Zavitsanos, Georgios Paliouras, George A. Vouros, and Sergios Petridis. Discovering subsumption hierarchies of ontology concepts from text corpora. In *WI '07: Proceedings of the IEEE/WIC/ACM International Conference on Web Intelligence*, pages 402–408, Washington, DC, USA, 2007. IEEE Computer Society.
- [4] B. Fortuna, D. Mladenic, and M. Grobelnik. Visualization of text document corpus. *Informatica Journal*, 29(4):497–502, 2005.
- [5] G. Salton, A. Wong, and C. S. Yang. A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620, 1975.
- [6] Philipp Cimiano, Andreas Hotho, and Steffen Staab. Learning concept hierarchies from text corpora using formal concept analysis. *Journal of Artificial Intelligence research*, 24:305–339, 2005.
- [7] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, 2003.
- [8] Jorge Morato, Miguel Ángel Marzal, Juan Lloréns, and José Moreiro. Wordnet applications. In *In: Proc. of the 2nd Int. Conf. Global WordNet*, 2004.